CLAIMS

1. A method comprising:

extracting DNA from a collection of samples from an organism;

preparing a sequence library with duplex adapters, wherein the sequence library is prepared by ligating a duplex adapter having a Unique Molecule Identifier (UMI) to an end of each of a plurality of strands of the extracted DNA and amplifying the extracted DNA with a first PCR;

selecting a subset of the sequence library;

amplifying the subset with a second PCR to increase a number of PCR duplicates;

sequencing a plurality of duplex reads from the amplified subset;

aligning the plurality of duplex reads to a host genome and denoising the plurality of duplex reads based on said alignment;

detecting the presence of a variant in at least one of the plurality of duplex reads; determining a signature of the variant;

comparing the signature of the variant to a collection of disease-specific variant signatures; and

determining a disease type based on the comparison.

2. The method of claim 1, wherein the UMI has less than five base pairs, optionally wherein the UMI has exactly three base pairs.

3. The method of claim 1, wherein the disease type is a cancer type, optionally wherein the cancer type comprises bladder cancer.

4. The method of claim 1, wherein the sequence library comprises one of a whole genome library or a whole exome library, optionally wherein preparing the sequence library with duplex adapters further comprises collapsing one or more errors on a strand of extracted DNA.

5. The method of claim 1, wherein amplifying the extracted DNA with the first PCR comprises removing sequencing errors based on the presence of two or more molecules with the same UMI.

6. The method of claim 1, wherein sequencing the plurality of duplex reads comprises sequencing on a paired-end system; or sequencing the plurality of duplex reads comprises sequencing on a single-end system and, wherein aligning the plurality of duplex reads to the host genome comprises: obtaining a plurality of single-end DNA sequencing reads;

separating a top-mapping strand of the DNA from a bottom-mapping strand of DNA; performing error collapsing on each of the top-mapping strands and the bottom-mapping strands; reverting the bottom-mapping strands to top-mapping strands by re-grouping based on UMI; and performing error correcting between the top and bottom strands; or sequencing the plurality of duplex reads comprises sequencing on a single-end system and, wherein aligning sequences to a host genome comprises: obtaining a plurality of single-end DNA sequencing reads; creating a synthetic paired-end read; and performing error correcting on all strands.

7. The method of claim 1, wherein sequencing the plurality of duplex reads comprises sequencing a series of uncorrected reads belonging to a duplex family, optionally further comprising processing uncorrected reads belonging to a duplex family to measure a read specific feature, and optionally further comprising filtering the uncorrected reads based on the measured read specific feature.

8. The method of claim 1, further comprising trimming the sequence of reads from the amplified subset.

9. The method of claim 1, wherein comparing the extracted variant to a collection of cancer-specific variant signatures comprises: calculating a tumor fraction estimation of a duplex-corrected signature; plotting the tumor fraction estimation to create a duplex-corrected signature for the extracted variant; and matching the duplex-corrected signature to a reference signature, optionally wherein the signature comprises a relative proportion of a trinucleotide mutation.

10. The method of claim 3, further comprising: correcting for library and sequencing artifacts by a panel of cancer-free controls sequenced on the same system; and estimating a tumor fraction.

11. The method of claim 1, further comprising integrating a genome-wide mutation from the sequencing reads as a weighted sum of single-base substitution (SBS) reference mutational signatures. optionally wherein integrating the genome-wide mutation comprises: deconvolving SBS mutational signatures from plasma DNA mixtures using a non- negative maximum likelihood model; estimating a tumor fraction by taking a weight of a tumor-associated SBS signature and normalizing by a total number of mutations and depth of sequencing; and calculating a signature score to determine that the cancer-associated SCS explains an observed mutation profile.

12. The method of claim 1, further comprising discarding a variant with an allele frequency greater than 30% optionally further comprising: aggregating reads having a variant with allele frequency less than 30%; calculating a frequency of variants in the aggregated reads trinucleotide context; and comparing the calculated trinucleotide variant frequency with a number of reference frequencies for different biological processes.

13. The method of claim 1, wherein determining a disease status comprises: randomly changing a trinucleotide frequency of a reference signature from the collection; performing a non-negative maximum likelihood fit between the randomly-permutated trinucleotide frequency and a frequency of the signature; and scoring the fit below a disease-negative threshold.

14. The method of claim 1, wherein the DNA is genomic DNA or wherein the DNA is cell-free DNA (cfDNA).

15. The method of claim 1, wherein detecting the presence of the variant comprises: providing the plurality of duplex reads to a pretrained machine learning model; and receiving therefrom an indication of a base variant irrespective of comparative sequence length.

16. The method of claim 15, wherein:

(i) the pretrained machine learning model comprises an artificial neural network, optionally wherein the artificial neural network is one of a feedforward neural network, a radial basis function network, a self-organizing map, learning vector quantization, a recurrent neural network, a Hopfield network, a Boltzmann machine, an echo state network, long short term memory, a bi-directional recurrent neural network, a hierarchical recurrent neural network, a stochastic neural network, a modular neural network, an associative neural network, a deep neural network, a deep belief network, a convolutional neural networks, a convolutional deep belief network, a large memory storage and retrieval neural network, a deep Boltzmann machine, a deep stacking network, a tensor deep stacking network, a spike and slab restricted Boltzmann machine, a compound hierarchical-deep model, a deep coding network, a multilayer kernel machine, or a deep Q-network; or

(ii) the pretrained machine learning model comprises a trained classifier, optionally wherein the trained classifier is a random decision forest.

17. The method of claim 1, wherein the collection of samples comprises a collection of plasma samples.

18. The method of claim 1, wherein the sequence library comprises a whole genome sequence library.

19. A computer program product for reducing sequencing error rates, the computer program product comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor to cause the process to perform a method according to any one of claims 1-18.